Neural Prediction Errors Distinguish Perception and Misperception of Speech

Abbreviated title: Neural prediction errors determine misperception

Authors: Helen Blank[1,2]*, Marlene Spangenberg[1,3], Matthew H. Davis[1]

Affiliations:

[1]MRC Cognition and Brain Sciences Unit, University of Cambridge, 15 Chaucer Rd, Cambridge, CB2 7E, UK.

[2]Department of Systems Neuroscience, University Medical Center Hamburg-Eppendorf Martinistr. 52, 20248 Hamburg, Germany.

[3]Department of Experimental Psychology, University of Oxford, 15 Parks Road, Oxford OX1 3PH, UK.

*Correspondence to: Helen Blank, Department of Systems Neuroscience, University Medical Center Hamburg-Eppendorf Martinistr. 52, 20248 Hamburg, Germany, hblank@uke.de

Number of Pages: 48, Number of Figures: 6, Number of Tables: 5

Number of Words: Abstract: 247, Introduction: 648, Discussion: 1674

Conflict of Interest: The authors declare no competing financial interests.

Acknowledgements:

This work was funded by the UK Medical Research Council (RG91365/SUAG/008 to MHD) and the EU Horizon 2020 programme (703635, Marie Curie fellowship to HB). We thank Yaara Erez, Jenni Rodd, Ediz Sohoglu, and Arnold Ziesche for valuable comments on a previous version of this manuscript and Helen Lloyd and Steve Eldridge for their assistance in radiography.

Abstract:

Humans use prior expectations to improve perception, especially of sensory signals that are degraded or ambiguous. However, if sensory input deviates from prior expectations, correct perception depends on adjusting or rejecting prior expectations. Failure to adjust or reject the prior leads to perceptual illusions especially if there is partial overlap (hence partial mismatch) between expectations and input. With speech, “Slips of the ear” occur when expectations lead to misperception. For instance, a entomologist, might be more susceptible to hear "The ants are my friends" for "The answer, my friend" (in the Bob Dylan song "Blowing in the Wind"). Here, we contrast two mechanisms by which prior expectations may lead to misperception of degraded speech. Firstly, clear representations of the common sounds in the prior and input (i.e., expected sounds) may lead to incorrect confirmation of the prior. Secondly, insufficient representations of sounds that deviate between prior and input (i.e., prediction errors) could lead to deception. We used cross-modal predictions from written words that partially match degraded speech to compare neural responses when male and female human listeners were deceived into accepting the prior or correctly reject it. Combined behavioural and multivariate representational similarity analysis of functional magnetic resonance imaging data shows that veridical perception of degraded speech is signalled by representations of prediction error in the left superior temporal sulcus. Instead of using top-down processes to support perception of expected sensory input, our findings suggest that the strength of neural prediction error representations distinguishes correct perception and misperception.

Significance Statement

Misperceiving spoken words is an everyday experience with outcomes that range from shared amusement to serious miscommunication. For hearing-impaired individuals, frequent misperception can lead to social withdrawal and isolation with severe consequences for well-

being. In this work, we specify the neural mechanisms by which prior expectations – which are so often helpful for perception – can lead to misperception of degraded sensory signals. Most descriptive theories of illusory perception explain misperception as arising from a clear sensory representation of features or sounds that are in common between prior expectations and sensory input. Our work instead provides support for a complementary proposal; namely that misperception occurs when there is an insufficient sensory representations of the deviation between expectations and sensory signals.

Introduction

The underlying neural signals that distinguish veridical and illusory perception remain unspecified. Perceptual illusions occur if sensory input deviates from prior expectations and perceivers fail to adjust or reject priors (Fletcher and Frith, 2009). Misperception is especially pronounced if there is partial overlap (and hence partial mismatch) between prior expectations and sensory input.

There are two plausible neural mechanisms for generating perceptual illusions. Firstly, misperception could arise due to clearer representations of the expected elements of sensory signals (McClelland and Elman, 1986; Norris et al., 2000). An alternative, prediction error theory (Mumford, 1992; Rao and Ballard, 1999; Friston, 2005) proposes a complementary mechanism; misperception occurs when neural representations of sensory signals that deviate from prior expectations are absent. Both these neural implementations of Bayesian perceptual inference can equally simulate a reduction of univariate activity for anticipated sensory signals (see Aitchison and Lengyel, 2017, Blank & Davis, 2016): 1) clearer representations of expected stimuli would lead to reduced noise or competition from alternative interpretations or 2) "prediction error" representations would be reduced for expected input. Both these theories are supported by the routine observation that neural activity is reduced for repeated stimuli,

(repetition suppression, see Henson, 2003; Grill-Spector et al., 2006; Summerfield et al., 2008). While reduced activity is plausibly due to a change in prior expectations (i.e., repeated stimuli are expected), it is not established whether repetition suppression is linked to reduced noise or reduced prediction errors in neural representations. In this work we distinguish these two explanations using repetition-induced slips of the ear – i.e., misperception of spoken words (Bond, 1999).

We therefore sought to measure speech representations in the left posterior STS (pSTS). This region shows effects of prior written word presentations on neural representations for degraded spoken words (Blank and Davis, 2016). Other studies have similarly shown influences of prior knowledge on pSTS activity during audio-visual speech processing (lip-reading: Nath and Beauchamp, 2011; Blank and von Kriegstein, 2013) and due to perceptual learning (Kilian-Hutten et al., 2011; Sohoglu and Davis, 2016; Bonte et al., 2017). Furthermore, multivariate pattern analysis (MPVA) shows syllable identity can be decoded from fMRI responses in the pSTS (Formisano et al., 2008; Evans and Davis, 2015).

We used presentations of written text to manipulate prior knowledge (Sohoglu et al., 2014) and recorded perceptual and neural (fMRI) responses to degraded (vocoded) spoken words (Shannon et al., 1995) (Fig 1A). Written and spoken words were combined into: (1) Match trials (i.e., written and spoken words were identical, e.g., whip-whip); (2) Total Mismatch trials (written/spoken words were phonologically unrelated, e.g., pit-corn); or (3) Partial Mismatch trials (written/spoken words had different initial or final sounds, e.g., kip-pip, pick-pip). Partial Mismatch trials lead to frequent misperception since listeners' often report that the written and spoken words match (Sohoglu et al., 2014). On each trial, participants provide a 4-alternative button press to report whether or not the spoken word matched the previous written word (1="definitely same", 2="possibly same", 3="possibly different", 4="definitely different").

Partial Mismatch trials manipulated which speech sounds were in common with or deviated from prior expectations (Fig 1B, Table 1) so as to distinguish two mechanisms for combining prior expectations and sensory signals. Firstly, speech-sensitive brain regions could represent sounds that are common between input and prior expectation (Kok et al., 2012): Clear representations of common sounds lead to confirmation of the prior (misperception) and unclear representations of common sounds to rejection (correct perception). Secondly, the brain could represent unexpected sounds that deviate between input and prior (i.e., prediction error (Rao and Ballard, 1999; Blank and Davis, 2016)): Clear representations of deviating sounds (prediction errors) lead to rejection of the prior and unclear representations of deviating sounds to confirmation (misperception, Fig 1B). These two mechanisms make distinct predictions for which pairs of Partial Mismatch trials will evoke similar patterns of neural activity in speech responsive regions (Fig 1C)) which we test with multivariate fMRI.

Materials and Methods

Design

In order to investigate the influence of prior expectations on the perception of degraded speech, behavioural responses and BOLD signals were acquired in an event-related fMRI design. Prior expectations were provided by presenting written words before degraded spoken words. The paring of written and degraded spoken words was manipulated in three conditions. 1) In Match trials (e.g., kit – kit) written and spoken words were identical. 2) In Total Mismatch trials (e.g., kit – ball) the spoken word was phonologically unrelated to the written word. 3) In Partial Mismatch trials the spoken and written word were phonologically different at the end of the word (Offset Mismatch; e.g., kit - kick) or were phonologically different at the beginning of the word (Onset Mismatch; e.g., kit - pit). Each condition contained 32 different word pairs that were repeated throughout the experiment. Behavioural responses were collected in a 4-

alternative-forced-choice task in which participants had to indicate whether they believed that the spoken word matched the previous written word (1 = “definitely same”, 2 = “possibly same”, 3 = “possibly different”, 4 = “definitely different”). In all following analyses, we merged responses 1 and 2 to “same” and 3 and 4 to “different” without considering confidence.

Ethics Statement

Ethical approval was provided by Cambridge Psychology Research Ethics committee (CPREC) under approval number 2009.46. All participants provided their written informed consent.

Participants

27 healthy native-English speakers (18-37 years) took part in the experiment after giving their informed consent. All participants were right-handed and reported normal or corrected-to-normal vision and no history of language, reading, or hearing impairments. Data from three participants had to be excluded (one due to technical problems during scanning, one due to an excessive number of missing behavioural responses (203 missed responses out of 1280, 15.86% missed responses) which was more than 4 SDs above the mean number of missed responses (M = 29.56, SD = 40.53)), and one because of aberrant behavioural responses (too few ”definitely different responses” in the Total Mismatch condition). The following analyses were therefore carried out using data from 24 participants (M = 24.17, SD = 5.01; 9 males and 15 females).

Stimuli

Stimuli consisted of 32 monosyllabic words, which were presented in spoken and written format. Auditory words were spoken by a male speaker of southern British English and recorded at 16-bit with a sampling rate of 44.1 kHz. The duration of spoken words ranged from

432 ms to 701 ms (M = 532, SD =64). The 32 words consisted of two sets of 16 words; each set containing a different vowel and items formed from four different onset and four different offset sounds (set 1: kit, kitsch, kip, kick, pit, pitch, pip, pick, writ, rich, rip, rick, wit, witch, whip, wick; set 2: corn, call, court, cork, torn, tall, taught, talk, born, ball, bought, baulk, warn, wall, wart, walk). Written and spoken words were combined in three conditions: 1) 32 Match pairs (identical written and spoken words, e.g., whip–whip), 2) 32 Partial Mismatch Onset and 32 Partial Mismatch Offset pairs (e.g., pit–kit, or pit-pitch), and 3) 32 Total Mismatch pairs (e.g., pit–corn). We selected item pairs in the Partial Mismatch trials carefully, so that we could group these pairs into quadruples with the same common sounds and deviating sounds between written and spoken forms. These common sound and deviating sound groups allow us to address our central research question concerning neural representations underlying speech perception and misperception (see Table 1 for a full list of item pairs and associated groups).

The amount of spectro-temporal detail of each spoken word was reduced by applying a noise-vocoding procedure (Shannon et al., 1995) using a custom-made Matlab (MathWorks Inc.) script. The script used 6 spectral channels that were logarithmically spaced between 70 and 5,000 Hz and superimposed the slow temporal envelope (low-pass filtered at 30Hz) onto corresponding band-pass filtered white noises. These parameters were chosen on the basis of previous perceptual data suggesting that they would result in high accuracy for Match and Total Mismatch trials and variable responses on Partial Mismatch trials ((Sohoglu et al., 2014), Expt 3).

Stimuli were delivered and behavioural responses recorded using E-Prime 2.0 software (Psychology Software Tools, Inc.). Visual stimuli were presented on a screen at the end of the scanner table, which participants could see through a mirror attached to the head-coil above their eyes. Auditory stimuli were presented binaurally through in-ear headphones (Sensimetrics

Corporation, Malden, MA, USA, model S14) after preprocessing to ensure a flat-frequency response and presentation at a comfortable listening volume.

Prior to scanning, participants completed two practice sessions. The first session was to familiarise participants with noise-vocoded speech and the "same/different" task to be used in the scanner. The second practice session was identical in task and timing to the main experiment and participants were given feedback and repeated practice to ensure that they made their responses within a 2.5s time-limit.

fMRI Procedure

The fMRI experiment lasted 75 minutes (5 MRI scanning sessions of 15 minutes). Each session included 300 randomised trials (256 event trials plus 44 null-events). We used a fast sparse-imaging protocol in which the duration of each trial was 3 seconds and noise-vocoded spoken words were presented in the silent gap between scans (see Fig 1A). Within each trial, a fixation cross was presented for 500 ms, followed by the written word presentation for 500 ms, and finally a window of 500 ms with the spoken word. This 500 ms delay between written and spoken word onset has been shown to be sufficient time to generate a prior expectation for the subsequent word (Sohoglu et al., 2014). During and after each vocoded word, a blank screen was presented for 1.5 seconds. Participants were given 2.5 seconds from the onset of each spoken word to make a 4-alternative response indicating whether the spoken word matched the preceding written word. Participants gave responses by pressing one of four buttons on a response box using the fingers of their right hand. Throughout the experiment, the words "same" and "different" were presented at the bottom of the screen to remind participants of the corresponding response buttons.

Each word occurred as a prior written word or spoken word with equal probability and each word pair was repeated twice within each scanning session so that there were ten presentations of each written-spoken word pair during the experiment. In addition to these experimental trials, each scanning session included 44 null events (trials without presentation of a written or spoken word) to aid estimation of a resting baseline.

Scanning Parameters

Structural Scanning

MRI data were acquired on a 3-Tesla Siemens Prisma scanner using a 32-channel head coil. A T1-weighted structural scan was acquired for each subject using a three dimensional MPRAGE sequence (TR 2250 ms, TE: 3.02 ms, flip angle: 9 deg, spatial resolution, 1x1x1 mm).

Functional Scanning

For each participant and scanning run, 312 echo planar imaging (EPI) volumes comprising 32 slices of 3 mm thickness were acquired using a continuous, descending acquisition sequence (TR 3000 ms, TA 2000 ms, TE 30 ms, FA 84 deg, matrix size: 64 x 64, in plane resolution: 3x3 mm, inter-slice gap 25%). Of these images, the first three EPI volumes were discarded (to allow for T1 equilibrium effects) and an additional nine EPI volumes were acquired after the last event of each scanning run. We used transverse-oblique acquisition, with slices angled away from the eyes.

Acoustic Similarity Analysis

Acoustic dissimilarity between spoken words was computed using methods described by (Billig et al., 2013). The matrix in Fig 2B illustrates the spectro-temporal similarity between stimuli. For each token, a Gammatone-based Fourier transform was computed, approximating

the frequency analysis performed by the ear. A spectral similarity matrix was then generated for each pair of tokens by comparing the spectral profile (on a log scale) of all time slices. Next, the maximum-similarity path through this similarity matrix was found using dynamic time warping. Summed similarity values along this path were computed and rank transformed such that the two most similar sound files were assigned a score of 0 and the two most dissimilar sound files were given a score of 1. As in Billig et al (2013), overall similarity reflects both shared vowels and consonants though vowel similarity has a greater influence. The spectral analysis and dynamic time warping were implemented in Matlab using existing functions for Gammatone spectral analysis and Dynamic Time Warping supplied by Ellis, available at http://www.ee.columbia.edu/~dpwe/resources/matlab/gammatonegram/ and http://labrosa.ee.columbia.edu/matlab/dtw.

Behavioural Analysis

First, we tested whether participants perceived word pairs in the Match condition as being the "same" and pairs in the Total Mismatch condition as being "different" with repeated-measures ANOVAs and paired t-tests (in Matlab, The MathWorks, Inc.)

Second, we tested perception in the Partial Mismatch condition. To determine whether the rate of misperception for individual items was due to sounds that were in common with, or deviated between the prior and input, we compared p("different") for each Partial Mismatch pair with two groups of word pairs. These groups either had the same sounds in common (common sound groups) or had the same deviating sounds (deviating sound groups). The goal of this analysis was to determine whether perceptual outcomes (i.e., responding "same" or "different") for a specific Partial Mismatch word pair (e.g., kit-pit) was better predicted by perception of: (1) three word pairs sharing the same deviating sounds (changing /k/ to /p/ as in kitsch-pitch, kip-

pip, and kick-pick) or (2) three word pairs sharing the same common sounds (common sounds /It/, as in pit-kit, writ-wit, wit-writ). To measure the similarity of behaviour in each of these groups we computed the sum squared difference between p("different") for each item pair with the mean of p("different") for three word pairs from the common sound group or three word pairs selected from the deviating sound group. The sum squared difference values were averaged over all Partial Mismatch items in each participant and over all participants for each item and entered into paired t-tests and ANOVAs by participants and items.

Univariate fMRI Analysis

Data were analysed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm) applying automatic analysis (aa) pipelines (Cusack et al., 2015). The first three volumes of each run were removed and the remaining scans were realigned to the first EPI image for each participant. The structural image was coregistered to the mean functional image and the parameters from the segmentation of the structural image were used to normalise the functional images, which were resampled to 2 mm isotropic voxels. The realigned normalised images were then smoothed with a Gaussian kernel of 8 mm full width half maximum. Data were analysed using a general linear model with a 128 s high pass filter. We included the onset of 7 event types in the GLM each convolved with the canonical SPM haemodynamic response: 7 conditions come from specifying the onset of spoken words paired with four types of written text, depending on perception: 1. Match perceived as "same", 2. Onset Partial Mismatch perceived as "same", 3. Offset Partial Mismatch perceived as "same", 4. Onset Partial Mismatch perceived as "different", 5. Offset Partial Mismatch perceived as "different", 6. Total Mismatch perceived as "different", and 7. Errors (i.e., Match perceived as "different" and Total Mismatch perceived as "same").

Following parameter estimation of the first level model, we conducted t-tests of 1) Total Mismatch perceived as "different" vs. Match perceived as "same" and 2) Partial Mismatch perceived as "different" vs. Partial Mismatch perceived as "same".

Multivariate fMRI Analysis

In the univariate analysis, we modelled BOLD responses combined over all item pairs with an individual condition (i.e., Match, Partial Mismatch Onset/Offset, Total Mismatch) but separated trials based on participants' behavioural responses ("same" vs "different"). This allows us to measure the impact of perception on the magnitude of neural responses in Partial Mismatch trials. In the multivariate analysis, the first level model was specified based on separating specific item pairs within each of the experimental conditions irrespective of behavioural responses (i.e., "same" and "different" responses were combined). This change was motivated for two reasons: Firstly, we wanted to avoid empty cells for single item pairs. This was necessary since for some participants there were word pairs that were always perceived as "same" or as "different" in all 10 repetitions of a particular Partial Mismatch trial. Secondly, we wanted to ensure that there were the same number of trials for each item pair included in the analysis. This avoids differences between neural representations for specific item pairs due to combining a different number of trials in the analysis.

Multivariate analyses were conducted on realigned data within each participant's native space without normalisation or spatial smoothing. An additional first level model was constructed for each participant. This model contained four conditions for which there were sufficient numbers of repetitions for item-specific modelling (Match, Total Mismatch, Onset Partial Mismatch, and Offset Partial Mismatch). Importantly, regressors for the 32 individual spoken words were used in each of these four conditions. This resulted in 128 conditions per participant per run.

For each of the 128 item-specific regressors we estimated single-subject T-statistic images for the contrast of speech onset compared to the unmodelled resting period, averaged over the five scanning runs.

We used the resulting single condition and item T-images (contrasted with the unmodelled resting baseline) for Representational Similarity Analysis (RSA) (Kriegeskorte et al., 2008) using the RSA toolbox (Nili et al., 2014).We used T-images so that effect sizes were weighted by their error variance which reduced the influence of large, but variable response estimates for multivariate analyses (Misaki et al., 2010). RSA involves testing whether the observed similarity of brain responses in specific conditions (a neural representational dissimilarity matrix or RDM) corresponds to a hypothetical pattern of similarity between these conditions (hypothesis RDM).

We constructed two hypothesis RDMs to test for greater similarity between word pairs. The first RDM tested word pairs that shared the common sounds between prior and spoken word in either Onset (e.g., **ki**t-**ki**tsch, **ki**p-**ki**ck; here the onset 'ki' is the same for both word pairs) or Offset (e.g., k**it**-p**it**, wr**it**-w**it**; here the offset 'it' is the same for both word pairs). The second RDM tested word pairs that shared the same deviating sounds between prior and spoken word in either Onset (e.g., **k**it-**p**it, **k**itch-**p**itch; here the different onsets '-k+p' are the same across word pairs) or Offset (e.g., ki**t**-ki**tsch**, pi**t**-pi**tch**; here the different offsets '-t+tʃ are the same across word pairs). Onset and Offset groups were combined in one single hypothesis RDM. We excluded between vowel comparisons to ensure that the results are not influenced by vowel representations which we have observed in previous studies (Evans and Davis, 2015; Blank and Davis, 2016). In addition, similarity between identical items (i.e., the main diagonal) was not included in our hypothesis RDMs (see Fig 5 C/D).

In a first step, we used these RDMs to test for differences between common and deviating sound groups without taking behaviour into account. In a second step, to determine whether representations of common or deviating sounds in the STS better explain perception and misperception of specific word-pairs, we used behavioural measures as weights in the RSA. Specifically, we averaged the rate of "different" responses across the four word pairs contributing to each common or deviating sound group, and rank ordered these groups in terms of the rate of accurate perception/misperception. With these ranks we constructed hypothesis RDMs for individual participants to test for similarity between word pairs that shared (1) common sounds in Partial Mismatch pairs or (2) deviating sounds in Partial Mismatch pairs while incorporating variability in perceptual outcomes. Our reasoning was that neural representations of common sounds in Partial Mismatch trials should be more apparent the more often a word pair is perceived as the "same" (see Fig 5A). Conversely, neural representations of deviating sounds should be stronger or more reliable for Partial Mismatch word pairs that are more often perceived as "different" (see Fig 5B for an illustration of these predictions). Since the weights in the hypothesis RDMs express expected dissimilarity values (i.e., higher values for higher dissimilarity which is the same as lower similarity), we reversed the ranking of the behavioural measures. For these analyses, we used perceptual outcomes for individual participants. Since we only aimed at testing for a monotonic relationship between perception and neural similarity, we rank-order behavioural response and used a Spearman correlation to test the relationship between hypothetical and neural RDMs. We rank transformed the proportion of "different" responses for Onset and Offset Partial Mismatch groups separately for each of the two vowel sets (/I/ and /ɔː/ as in "kick" and "tall"). This ensured that these analyses link the rate of perception/misperception to informative neural representations of common/deviating sounds rather than to differences in the representation of the two vowels in

our stimulus set (since these gave rise to different overall rates of speech perception/misperception).

Region of Interest (ROI) Definition

Our key question concerned neural representation of Partial Mismatch trials in the left posterior STS; a region previously shown to integrate prior expectations and spoken words (Blank and Davis, 2016). Importantly for our RSA analysis approach, multivariate BOLD signals have been used to decode syllable identity in several previous studies (Formisano et al., 2008; Boets et al., 2013; Du et al., 2014; Evans and Davis, 2015; Blank and Davis, 2016). To locate this ROI for multivoxel representational similarity analysis (RSA), we compared neural responses to Total Mismatch and Match trials (t-contrast: Total Mismatch ("different" response) > Match ("same" response) at $p < 0.001$). In addition, to remove activations that extended into adjacent parietal regions, we applied a mask of the combined STG and MTG clusters from the Harvard-Oxford Cortical Structural Atlas. The size of the ROI was a volume of 1148 mm3 corresponding to 34 voxels in the RSA voxel size if 3x3x3.75. The same ROI was used for the analysis of deviating and common sound groups. This ROI definition is based on entirely independent conditions (Total Mismatch vs. Match) from the conditions used in the RSA analysis (which is focussed on Partial Mismatch trials). Furthermore, this ROI definition does not favour the representation of either deviating or common sounds in the main RSA analysis. Our previous work (Blank and Davis, 2016) has shown that univariate activation differences between unexpected and expected stimuli are equally consistent with two types of neural computation: A sharpening model (without representation of prediction errors) explains the decreased response in the Match condition as due to a suppressed representation of unexpected features; i.e., a reduced representation of deviating sounds and an enhanced representation of common sounds. Alternatively, a Prediction Error model explains the decreased response in

the Match condition as due to reduced prediction errors which reduce the representation of common sounds and enhanced representations of deviating sounds in Total Mismatch conditions (Blank and Davis, 2016).

Searchlight Analysis

We conducted a whole brain searchlight analysis, to make sure that we do not overlook significant effects outside of the ROI we defined a-priori. We measured multivoxel neural RDMs by computing the dissimilarity (1 - Pearson correlation across voxels) of T-statistics for all possible combinations of items and conditions. In a searchlight analysis, the sets of voxels were extracted by specifying grey-matter voxels (voxels with a value > 0.20 in a probabilistic grey-matter map) within a 10 mm radius sphere of each grey matter voxel (with a voxel size of 3x3x3.75 mm, i.e., a maximum of 65 voxels per sphere). This was repeated for all searchlight locations in the brain. The similarity between the observed RDM, and each of the hypothetical RDMs was computed using a Spearman correlation for each searchlight location and the resulting correlation coefficient returned to the voxel at the centre of the searchlight. This resulted in a Spearman correlation map for each participant in each grey matter voxel. To assess searchlight similarity values across participants at the second level, the Spearman correlation maps for each participant were Fisher-z-transformed to conform to Gaussian assumptions, normalized to MNI space, and spatially smoothed with a 10 mm FWHM Gaussian kernel for group analysis. For a visualization of our RSA procedure see Figure 2 in (Kriegeskorte et al., 2008). We extracted similarity values from searchlights within our ROI defined using the independent contrast from the univariate fMRI analysis.

Region of Interest (ROI) Analysis

In addition to using the ROI defined by the univariate analysis Total Mismatch ("different" response) > Match ("same" response) as a search volume in the whole brain RSA analysis (previous section), we used this ROI to extract neural RDMs from the Partial Mismatch conditions to test for representations of deviating and common sounds. Specifically, we correlated the neural RDM from this ROI with the behaviourally weighted hypothesis RDMs for deviating and common groups. We conducted one-sample t-tests on the obtained a Fisher-z-transformed Spearman correlation value for these two RDMs to test whether the correlation was significantly greater than zero for the two conditions, individually. We then tested for differences between these conditions in a paired t-test. This approach allows us to specifically test the representation in our a-priori defined ROI. There are some methodological differences between the whole brain searchlight and the ROI approach, such as 1) the same number of vowels per sphere across all searchlight locations across the brain vs. one fixed cluster size in the ROI approach, 2) grey matter masking in the searchlight approach (none in the ROI approach), 3) comparison of searchlight locations across subjects in MNI space vs. transformation of individual ROIs to subjects' native space in the ROI approach.

Results

Partial mismatch with prior expectations leads to frequent misperception

Behavioural responses confirmed that participants correctly perceived written and spoken word pairs in the Match condition as identical and pairs in the Total Mismatch condition as "different". Perception in the Partial Mismatch condition was more variable such that listeners were often deceived into reporting that spoken word matched the written prior (Fig 2C-E). A repeated-measures one-way analysis of variance revealed significant differences between these three conditions ($F(23,2) = 603.303$, $p < 0.001$). Post-hoc paired t-tests confirmed more "same"

responses in Match than Partial Mismatch conditions ($t(23) = 17.719$, $p < .001$), and in Partial Mismatch than in Total Mismatch conditions ($t(23) = 11.782$, $p < .001$).

Within the Partial Mismatch trials, the rate of "different" responses was related to a measure of acoustic similarity/dissimilarity between expected and heard speech. Acoustic dissimilarity between 1) 6-channel vocoded spoken words and 2) between 6-channel vocoded and clear spoken words was computed using correlation methods described by (Billig et al., 2013). Degraded spoken words that were more similar to the 6-channel vocoded acoustic form of the preceding written word were more often judged to be identical and more dissimilar spoken word pairs were more often judged to be "different" ($r(62) = 0.3906$, $p = 0.0014$, Fig 2 B/C). However, this correlation with behaviour in the Partial Mismatch condition was not apparent for similarity between 6-channel vocoded and clear spoken words ($r(62) = -0.0039$, $p = 0.9754$), but only when all conditions including Match and Total Mismatch were considered ($r(126) = 0.5195$, $p < 0.001$ for clear-to-degraded similarity, and $r(126) = 0.8859$, < 0.001 for degraded-to-degraded similarity). However, this finding does not explain whether it is the acoustic similarity of common sounds or acoustic dissimilarity of deviating sounds that is more important for determining perception and misperception in Partial Mismatch trials. To explore this issue we use between-item and between-participant variation in perception of Partial Mismatch trials (depicted in Fig 2D).

To determine whether perception depends more on common or deviating sounds between prior written text and degraded speech input we compared rates of perception and misperception for each Partial Mismatch word pair with two other groups of word pairs (Fig 3A). This analysis assessed whether perception of a specific Partial Mismatch word pair (e.g., kit-pit) is better predicted by perception of three other word pairs that share: (1) the same common sounds (i.e., p**it**-k**it**, w**it**-wr**it**, and wr**it**-w**it,** which all contain a common offset /**It**/) or (2) the same deviating

sounds (i.e., **k**ip-**p**ip, **k**itsch-**p**itch, **k**ick-**p**ick**,** which all contain a deviating onset /**k**/ and /**p**ʃ/). All 64 Partial Mismatch item pairs (32 onset, and 32 offset mismatch pairs) were grouped into 16 common sound groups and 16 deviating sound groups (full list in Table 1). Within each group we computed the sum square difference of response rates (i.e., proportion of ”different” responses) so as to assess whether more consistent behavioural responses were apparent for Partial Mismatch word pairs grouped by their common or deviating sounds.

Responses to Partial Mismatch pairs were significantly more similar (i.e., lower sum square difference) for word pairs sharing the same deviating sounds than for items sharing the same common sounds (paired t-tests over items: $t(63) = 6.744$, $p < 0.001$ and participants: $t(23) = 10.567$, $p < 0.001$, averaged data shown in Fig 3B). Behavioural performance is more homogenous when different Partial Mismatch item pairs are grouped according to the deviating sound, as compared to groups organized according to the common sound. These results therefore indicate that speech perception and misperception are better predicted by the specific speech sounds that deviate from prior expectation than by the sounds that are consistent with prior expectations.

For completeness, we ran additional exploratory analyses on behavioural data separating Partial Mismatch trials with different vowels and Onset/Offset mismatch. For p(“different”, Fig 2F) ANOVAs by participants (F1) and items (F2) showed significant main effects of vowel identity ($F1(1,23) = 117.413$, $p < 0.001$; $F2(1,60) = 20.64$, $p < 0.001$) and Onset/Offset ($F1(1,23) = 13.925$, $p = 0.001$; though this was only a trend by items: $F2(1,60) = 3.37$, $p = 0.0712$) as well as an interaction ($F1(1,23) = 51.219$, $p < 0.001$; $F2(1,60) = 5.42$, $p = 0.0233$).

In addition, we conducted ANOVAs on Sum Squared Difference values derived from the behavioural data (Fig 3C). For word pairs grouped by deviating sounds, there was no main effect of vowel ($F1(1,23) = 0.152$, $p = 0.6999$; $F2(1,60) = 0.01$, $p = 0.9133$), and no consistent

effect of Onset and Offset (F1(1,23) = 11.106, p = 0.0029; F2(1,60) = 1.78, p = 0.1876) or interaction of vowel and Onset/Offset (F1(1, 23) = 56.164, p < 0.001; F2(1,60) = 1.76, p = 0.1892). For the common sound groups, there were no significant effects (main effect of vowel: F1(1,23) = 0.375, p = 0.5462; F2(1,60) = 0.08, p = 0.7839; main effect of Onset and Offset: F1(1,23) = 2.361, p = 0.1380; F2(1,60) = 0.6, p = 0.4414; and interaction of vowel and Onset/Offset: F1(1, 23) = 0.683, p = 0.4169; F(1,60) = 0.13, p = 0.7203). Given the lack of significant effects in item analyses and our between-item manipulation of vowel and Onset/Offset mismatch, findings from the analysis across participants are potentially false-positives. We did not have specific hypotheses regarding the influence of these other factors and hence further studies are needed to follow up on how vowel identity and position of mismatch influence perception and neural representations.

Univariate magnitude of BOLD activity increases during perception of mismatch

Next, we analysed fMRI responses to assess how the magnitude of neural responses differed between trials in which matching and mismatching text preceded spoken words. We replicated previous results (Sohoglu et al., 2012; Blank and Davis, 2016) showing significantly greater activity for Total Mismatch than Match trials in the bilateral superior temporal sulcus (STS, p < 0.05 FWE-corrected, Fig 4, Table 2). We further showed that the magnitude of the BOLD signal was increased for Partial Mismatch pairs heard as "different" compared to the same word pairs heard as "same" in a largely overlapping brain network including the left STS (Fig 4, Table 3). Brain regions in and around the left posterior STS have long been known to support perceptual processing of speech (Scott and Johnsrude, 2003; Hickok and Poeppel, 2007) and to integrate expectations from different modalities with speech input (Noppeney et al., 2008; Sohoglu et al., 2012; Blank and von Kriegstein, 2013; Blank and Davis, 2016).

In addition, we examined the magnitude of the univariate activity in the overlapping left pSTS region identified using Total Mismatch ("different" response) > Match ("same" response) and Partial Mismatch ("different" response) > Partial Mismatch ("same" response) (Fig 4B). We did neither find a significant difference between Total Mismatch ("different" response) and Partial Mismatch ("different" responses): $t(23) = 1.6172$, $p = 0.1195$, nor between Match ("same" response) and Partial Mismatch ("same" responses): $t(23) = 0.9782$, $p = 0.3381$. We also observed a difference in univariate activation in the left postcentral gyrus. This is plausibly due to differential difficulty of the button presses responses that participants made with the right hand and need not reflect a speech-specific process. However, since the superior temporal sulcus has not been shown to process finger movements it seems implausible that a similar explanation could apply to differential activity for match and mismatch trials in the STS.

Neural representations of deviating, not common sounds are linked to (mis)perception

We used multivariate, representational similarity analysis (Kriegeskorte et al., 2008; Nili et al., 2014) to distinguish between representations of deviating and common sounds in Partial Mismatch trials. We defined an independent STS region of interest (ROI based on the contrast of Total Mismatch > Match trials, at $p < 0.001$ uncorrected, inclusively masked with Superior and Middle Temporal Gyrus regions from the Oxford-Harvard Atlas (Desikan et al., 2006)). In this search volume, we first test for similarity between Partial Mismatch word pairs that shared (1) common sounds between prior and spoken word at syllable onset (e.g., **ki**t-**ki**tsch, **ki**p-**ki**ck) and offset (e.g., k**it**-p**it**, wr**it**-w**it)** or (2) deviating sounds between prior and spoken word at syllable onset (e.g., **k**it-**p**it, **k**ip-**p**ip) and offset (e.g., ki**p**-ki**ck**, pi**p**-pi**ck**, Table 1). These analyses showed a significant representation of deviating sounds for searchlight locations in our STS ROI ($x = -63$, $y = -40$, $z = 9$, pFWE (small volume corrected (svc)) = 0.017, $t(23) = 3.18$) and a marginally significant trend for representations of common sounds in the same

region (x = -66, y = -34, z = 12, pFWE(svc) = 0.059, t(23) = 2.58). However, a paired t-test revealed no significant difference between these representations (pFWE(svc) = 0.656, t(23) = 0.41). This analysis provides some limited evidence for neural representations of deviating sounds in Partial Mismatch trials and is equivocal concerning representations of common sounds. Hence, the results provide no clear evidence to favour one or other type of neural representation.

To determine whether representations of deviating or common sounds in the STS better explain perceptual outcomes, we conducted a further multivariate analysis which included participant-specific measures of the rate of perception and misperception for common and deviating sound groups (Table 1). To do this, we averaged the rate of "different" responses across the four word pairs contributing to each common or deviating sound group, and rank ordered these groups in terms of the rate of accurate perception or misperception. If representations of common sounds in Partial Mismatch trials determine perception, then stronger representations of these common sounds should correlate with more frequent "same" responses (i.e., misperception, see Fig 5A). Conversely, if representations of deviating sounds determine perception, then stronger representations of these sounds should be apparent for Partial Mismatch pairs that are more often perceived as "different" (i.e., correct perception, see Fig 5B). For this analysis, we used behavioural measures from individual participants, rank transformed separately for each of the two vowel sets (/I/ and /ɔ:/ as in "kick" and "tall") and for Onset/Offset Mismatch pairs (see Table 1). This ensured that these analyses exclude otherwise uninteresting differences between the rate of perception/misperception for the two vowels and Onset/Offset mismatches. By using rank correlations, we tested for any monotonic relationship between perceptual outcomes and neural representations without requiring a linear relationship.

In the searchlight analysis, we correlated neural RDMs from each searchlight sphere with two hypothesis RDMs containing behavioural responses as similarity weights for word pairs either grouped based on the 1) deviating sounds or 2) common sounds in the item pairs (schematically depicted in Fig 5 A and B). When we applied small volume correction for our STS search volume (Fig 5 E), there was a positive correlation between single subject measures of perception (i.e., "different" responses) with neural representations of deviating sounds (pFWE(svc) = 0.01, $t(23) = 3.46$, x = -66, y = -25, z = 5) and no correlation of misperception (i.e., "same" responses) with representations of common sounds (pFWE(svc) = 0.693, $t(23) = 0.28$). A paired t-test further showed that the correlation with deviating representations was more reliable than the correlation with representations of common sounds (deviating vs. common sound groups paired t-test: pFWE(svc) = 0.042, $t(23) = 2,74$, x = -66, y = -25, z = 5). To visualise the outcome of this analysis (Fig 5 F/G), we computed the average neural similarity among the four item pairs within each group for each participant and averaged the rank ordered item pairs over participants based on the proportion of "different" responses (i.e., as shown schematically for one mismatch position and vowel in Fig 5 A/B).

We supplemented this searchlight analysis, by extracting a Fisher-z-transformed Spearman correlation value for each of the two analyses with a pattern similarity computed for the whole of the ROI. BOLD pattern similarity computed over all voxels in the ROI correlated with individual participants' rates of "different" responses for word pairs grouped according to deviating sounds (r = 0.0858, one-sample t-test: $t(23) = 2.5715$, $p = 0.0171$). Furthermore, the equivalent correlation was non-significant for common representations; higher rates of responding "same" were not correlated with representational similarity for words pairs grouped according to common sounds ($r = -0.0253$, one-sample t-test: $t(23) = -0.7946$, $p = 0.4350$). Again, a comparison of these two correlations with a paired t-test showed significantly more

reliable correlations between perceptual outcomes with prediction error representations than with expected representations ($t(23) = 2.6472$, $p = 0.0144$).

To ensure that effects in other brain areas were not missed, we also inspected whole brain searchlight results for these three multivariate analyses (Fig 6). This did not reveal any further areas that reached a whole-brain corrected threshold, but showed two clusters in left motor and frontal regions at $p < 0.001$ uncorrected for the paired t-test comparing deviating vs. common sound groups (see Table 4). The left motor cluster was also observed for the correlation between behavioural responses and representations of deviating sounds in Partial Mismatch trials (Table 5). No searchlight locations reached $p < 0.001$ uncorrected for correlation with representation of common sounds.

In a further exploratory analysis and to generate hypothesis for future studies, we also examined neural representations in another cluster that showed activation differences in the univariate analysis. Specifically, we examined the cluster in the left middle frontal gyrus revealed by the independent univariate contrast 'Total Mismatch ("different" response) > Match ("same" response)' (peak at x = -32, y = 20, z = 32) for small volume correction, because this region has previously been reported to contain representations of prior information during perception of degraded speech (Blank and Davis, 2016; Sohoglu and Davis, 2016; Cope et al., 2017). Here, we found a significant representation of deviating sound groups (x = -33, y = 11, z = 31, pFWE(svc) = 0.004), no representation of common sound groups (x = -27, y = 26, z = 35, pFWE(svc) = 0.631). This difference was also significant in a paired t-test x = -39, y = 17, z = 28, pFWE(svc) = 0.007).

Discussion

Misperceiving spoken words is a common, everyday experience with outcomes that range from shared amusement to serious miscommunication. For hearing-impaired individuals, frequent

misperception can lead to social withdrawal and isolation with severe consequences for well-being (Dalton et al., 2003). In this work, we specify the neural mechanisms by which prior expectations – which are so often helpful for perception – can lead to deception when perceiving degraded sensory signals.

We induced frequent misperception of speech by providing clear prior expectations (written text) that partially matched/mismatched with degraded spoken words. Listeners often reported that a spoken word with one mismatching sound was the same as previously presented text (e.g., reporting that pairs like pick–kick, or pick-pip are the "same"). Behavioural results revealed that perceptual outcomes for these pairs were more similar to perceptual outcome for other word pairs that shared the same deviating sounds (i.e., "-p.. +k.." or "-..k +..p", in the examples above) than for word pairs that shared the common sounds (i.e., ".ick", or "pi."). However, this behavioural observation does not determine the underlying neural mechanisms that support perception and misperception of speech.

Our fMRI data showed reductions in the magnitude of the univariate BOLD signal in the left pSTS (Fig 4) for written/spoken word-pairs that are heard as "same". This effect does not seem to reflect passive adaptation since the magnitude of the reduction does not depend on the number of shared/deviating segments (i.e., Partial Mismatch and Total Mismatch trials respond similarly) but on the perceptual outcome (i.e., whether participants respond "same" or "different"). Thus, the influence of prior knowledge on lower-level speech processing is linked to trial-by-trial perceptual outcomes (i.e., detecting deviating sounds in Partial Mismatch pairs). However, these response reductions do not determine the neural mechanisms responsible (see (Blank and Davis, 2016; Aitchison and Lengyel, 2017)). Reduced univariate activity for matching trials could be due to, either: (1) more efficient / less effortful processing of common sounds (Murray et al., 2004; Kok et al., 2012; Blank and Davis, 2016) or (2) suppressed

processing of common sounds (i.e., explaining away) (Murray et al., 2004; Friston, 2005; Blank and Davis, 2016). Both of these proposals can explain reductions in the magnitude of neural responses for partially matching trials that are heard as "same" and other similar findings from repetition suppression designs. Hence, we used RSA fMRI to measure representational content in the pSTS so as to specify the neural mechanisms by which listeners combine prior knowledge and degraded sensory signals. Specifically, we can decode whether the repeated (i.e., expected) or the non-repeated (unexpected) part of the stimulus is preferentially represented in the pSTS and hence how representations of (un)expected elements of degraded stimuli are linked to perception.

The findings of our multivariate fMRI analyses confirm representations of prediction error in the STS. Neural representations of deviating sounds were correlated with perceptual outcomes, i.e., neural representations of prediction error were more apparent for trials in which written/spoken mismatch was detected. The equivalent correlation with perceptual outcomes for representations of expected sounds was non-significant (and showed a numerical trend in the non-predicted direction). Furthermore, there was a significant difference between the positive correlation for deviating sounds and the null correlation for common sounds. While our methods do not permit us to draw conclusions from the absence of a significant effect, we note that effect sizes for our reliable multivariate analyses are in line with those seen in previous, similar fMRI studies (Evans and Davis, 2015; Blank and Davis, 2016).

We therefore conclude that neural representations of prediction error are apparent in the pSTS and linked to perceptual outcomes during perception of degraded speech. These findings are best explained by the proposal that neural representations in the pSTS signal prediction error, i.e., representations of the speech sounds that deviate from prior expectations. These findings are well explained by accounts of speech perception which assign an important

role to predictive coding computations (Arnal et al., 2011; Giraud and Poeppel, 2012; Blank and Davis, 2016).

Our previous work also provided evidence for a predictive coding account of speech perception by showing (in the pSTS) an interaction such that increased sensory detail had opposite effects on multivariate speech representations following neutral and matching text (Blank and Davis, 2016). While differences in the neural representation of speech in this previous work could be due to changes in listening strategy – e.g., listeners anticipating that degraded speech will be harder to understand following neutral text, or being distracted by prior presentation of written text – these alternative explanations could not apply to the present study in which all spoken words were preceded by written text. The present study also goes beyond our previous work by directly linking perceptual outcomes to neural representation of prediction error. That is, trials that evoke clearer neural representations of deviating sounds (i.e., prediction errors) in the pSTS lead to more accurate perception.

Alternative theories of speech perception – most notably interactive activation accounts such as the TRACE model (McClelland and Elman, 1986) – have proposed that perception depends on joint activation of common representations between prior expectations and speech signals. Our experimental design allowed several tests for the representation of these common sounds during Partial Mismatch trials, but our neural data provides no evidence for neural representations of expected sounds as proposed by interactive activation models. Thus, instead of using top-down processes to support the perception of expected sounds (a mechanism which has previously been criticised as too vulnerable to hallucination (Norris et al., 2000)), we propose that neural representations of prediction error play a critical role in achieving accurate perception of speech. Listeners use representations of prediction error as a signal update or overrule prior expectations when these are incompatible with incoming signals. Stronger

prediction error signals therefore lead to correct rejection of prior expectations and more accurate perception of degraded speech. While our findings challenge interactive activations accounts of perception (McClelland and Elman, 1986), we cannot rule out some predictive coding theories (Rao and Ballard, 1999; Friston, 2005) in which representations of prediction error and expected sounds (i.e., top-down predictions) are computed in parallel in different sets of neurons or cortical laminae. It remains to be seen whether other methods (e.g., laminar-specific analysis of ultra-high field fMRI) can be used to demonstrate a representation of expected sounds that are detected in degraded signals, or whether expected sounds are not directly represented in neural responses.

One avenue for future investigation could be to explore other influences of prior knowledge in perception. For example, recent multivariate fMRI studies have shown changes to neural representations of ambiguous speech sounds due to adaptation or phonetic recalibration training (Kilian-Hutten et al., 2011; Bonte et al., 2017). These decoding techniques demonstrate that neural representations in the posterior STS can discriminate between different perceptual outcomes for ambiguous sounds due to learning. However, so far these findings do not reveal the mechanisms underlying these neural representations, i.e., they do not distinguish between sharpening and prediction error mechanisms (Kilian-Hutten et al., 2011; Bonte et al., 2017), although other studies have shown common neural changes due to prior knowledge and perceptual learning in line with predictive coding (Sohoglu and Davis, 2016). Future work could test of these claims using multivariate fMRI methods. Critically for computations of prediction error, correspondences between sensory signals and prior expectations can either enhance or suppress informative neural representations (depending on signal quality and perceptual outcomes, Blank and Davis, 2016)) whereas sharpening accounts

propose that neural representations of signals are always enhanced by accurate expectations. Further tests of these proposals in the context of perceptual learning would be informative.

In addition to the laboratory-induced occurrences of speech misperception that we have studied here, prediction error representations have the potential to explain more ecologically- and clinically-significant instances of misperception. For example, in naturally-occurring slips of the ears, listeners typically report incorrect, but phonological and lexically well-formed content words while adding or modifying function words to generate plausibly structured phrases and sentences (Bond, 2005). Thus, real-world misperception of speech involves both sensory confusions (i.e., content words are misidentified), and the filling in of predicted words. These observations seem to follow naturally from an account in which misperception derives from weak representations of prediction error. Older individuals have a double vulnerability to speech misperception; age-related hearing loss is the most common sensory impairment in old age (Roth et al., 2011) and even when intelligibility is equated older listeners are more likely than younger listeners to report a predictable, but incorrect word (Rogers and Wingfield, 2015). This is consistent with a novel proposal derived from the current study that impaired sensory processing in older listeners leads to a systematic reduction in the strength or efficacy of prediction error representations.

Our account of misperception based on inadequate prediction error representations is also relevant to abnormal perceptual experience arising from over-application of prior beliefs about the world without incoming sensory information (i.e., hallucinations). Inappropriate integration of prior expectations could lead to verbal hallucinations ranging from voice hearing in individuals without any clinical diagnosis (Alderson-Day et al., 2017) to the more distressing experiences reported by individuals with schizophrenia (Fletcher and Frith, 2009). Recent work has shown that individuals with early psychosis and healthy-individuals at risk of psychosis

show a greater reliance on prior knowledge during perception of visually-degraded images (Teufel et al., 2015). Our observations of neural representations that underpin prior knowledge induced misperceptions of speech may therefore assist in exploring the origins of auditory-verbal hallucinations in psychosis.

The present findings show that representations of prediction error determine perceptual outcomes in listening conditions that lead to frequent misperceptions. Most descriptive theories explain illusory perception as arising from sensory representations of features or sounds that are supported by prior expectations (Gregory, 1997). Our work instead provides support for a complementary proposal; namely that misperception occurs when there is an insufficient sensory representations of the difference between expectations and sensory signals. Sensory prostheses, or other neural interventions (Moore and Shannon, 2009; Zoefel and Davis, 2017) that enhance representations of prediction error may thereby improve the accuracy of speech perception in hearing impaired individuals.

References:

Aitchison L, Lengyel M (2017) With or without you: predictive coding and Bayesian inference in the brain. Current Opinion in Neurobiology 46:219–227.

Alderson-Day B, Lima CF, Evans S, Krishnan S, Shanmugalingam P, Fernyhough C, Scott SK (2017) Distinct processing of ambiguous speech in people with non-clinical auditory verbal hallucinations. Brain 140:2475–2489.

Arnal LH, Wyart V, Giraud A-L (2011) Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nature Neuroscience 14:797–801.

Billig AJ, Davis MH, Deeks JM, Monstrey J, Carlyon RP (2013) Lexical influences on auditory streaming. Current biology : CB 23:1585–1589.

Blank H, Davis MH (2016) Prediction Errors but Not Sharpened Signals Simulate Multivoxel fMRI Patterns during Speech Perception. PLOS Biology 14:e1002577.

Blank H, von Kriegstein K (2013) Mechanisms of enhancing visual–speech recognition by prior auditory information. NeuroImage 65:109–118.

Boets B, Op de Beeck HP, Vandermosten M, Scott SK, Gillebert CR, Mantini D, Bulthe J, Sunaert S, Wouters J, Ghesquiere P (2013) Intact But Less Accessible Phonetic Representations in Adults with Dyslexia. Science 342:1251–1254.

Bond ZS (1999) Slips of the ear : errors in the perception of casual conversation. San Diego (Calif.) : Academic press. Available at: http://lib.ugent.be/catalog/rug01:000912907.

Bond ZS (2005) Slips of the Ear. In: The Handbook of Speech Perception, pp 290–310. Blackwell Publishing Ltd. Available at: http://dx.doi.org/10.1002/9780470757024.ch12.

Bonte M, Correia JM, Keetels M, Vroomen J, Formisano E (2017) Reading-induced shifts of perceptual speech representations in auditory cortex. Scientific Reports 7:5143.

Cope TE, Sohoglu E, Sedley W, Patterson K, Jones PS, Wiggins J, Dawson C, Grube M, Carlyon RP, Griffiths TD, Davis MH, Rowe JB (2017) Evidence for causal top-down frontal contributions to predictive processes in speech perception. Nature Communications 8:2154.

Cusack R, Vicente-Grabovetsky A, Mitchell DJ, Wild CJ, Auer T, Linke A, Peelle JE (2015) Automatic analysis (aa): efficient neuroimaging workflows and parallel processing using Matlab and XML. Frontiers in Neuroinformatics 8 http://journal.frontiersin.org/Journal/10.3389/fninf.2014.00090/pdf.

Dalton DS, Cruickshanks KJ, Klein BEK, Klein R, Wiley TL, Nondahl DM (2003) The Impact of Hearing Loss on Quality of Life in Older Adults. The Gerontologist 43:661–668.

Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ (2006) An automated labeling

system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31:968–980.

Du Y, Buchsbaum BR, Grady CL, Alain C (2014) Noise differentially impacts phoneme representations in the auditory and speech motor systems. P Natl Acad Sci USA 111:7126–7131.

Evans S, Davis MH (2015) Hierarchical Organization of Auditory and Motor Representations in Speech Perception: Evidence from Searchlight Similarity Analysis. Cereb Cortex 25:4772–4788.

Fletcher PC, Frith CD (2009) Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. Nat Rev Neurosci 10:48–58.

Formisano E, De Martino F, Bonte M, Goebel R (2008) "Who" is saying "what"? Brain-based decoding of human voice and speech. Science 322:970–973.

Friston K (2005) A theory of cortical responses. Philos Trans R Soc Lond B Biol Sci 360:815–836.

Giraud AL, Poeppel D (2012) Cortical oscillations and speech processing: emerging computational principles and operations. Nat Neurosci 15:511–517.

Gregory, Richard (1997) Eye and Brain: The Psychology of Seeing. Oxford University Press.

Grill-Spector K, Henson R, Martin A (2006) Repetition and the brain: neural models of stimulus-specific effects. Trends Cogn Sci 10:14–23.

Henson RNA (2003) Neuroimaging studies of priming. Prog Neurobiol 70:53–81.

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8:393–402.

Kilian-Hutten N, Valente G, Vroomen J, Formisano E (2011) Auditory Cortex Encodes the Perceptual Interpretation of Ambiguous Sound. J Neurosci 31:1715–1720.

Kok P, Jehee JFM, de Lange FP (2012) Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. Neuron 75:265–270.

Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis - connecting the branches of systems neuroscience. Front Syst Neurosci 2:4.

McClelland JL, Elman JL (1986) The TRACE model of speech perception. Cognitive psychology 18:1–86.

Misaki M, Kim Y, Bandettini PA, Kriegeskorte N (2010) Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. NeuroImage 53:103–118.

Moore DR, Shannon RV (2009) Beyond cochlear implants: awakening the deafened brain. Nat Neurosci 12:686–691.

Mumford D (1992) On the Computational Architecture of the Neocortex .2. The Role of Corticocortical Loops. Biol Cybern 66:241–251.

Nath AR, Beauchamp MS (2011) Dynamic Changes in Superior Temporal Sulcus Connectivity during Perception of Noisy Audiovisual Speech. J Neurosci 31:1704–1714.

Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N (2014) A toolbox for representational similarity analysis. PLoS computational biology 10:e1003553.

Noppeney U, Josephs O, Hocking J, Price CJ, Friston KJ (2008) The effect of prior visual information on recognition of speech and sounds. Cereb Cortex 18:598–609.

Norris D, McQueen JM, Cutler A (2000) Merging information in speech recognition: Feedback is never necessary. Behav Brain Sci 23:299–+.

Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2:79–87.

Rogers CS, Wingfield A (2015) Stimulus-independent semantic bias misdirects word recognition in older adults. The Journal of the Acoustical Society of America 138:EL26-EL30.

Roth TN, Hanebuth D, Probst R (2011) Prevalence of age-related hearing loss in Europe: a review. European Archives of Oto-Rhino-Laryngology 268:1101–1107.

Scott SK, Johnsrude IS (2003) The neuroanatomical and functional organization of speech perception. Trends Neurosci 26:100–107.

Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M (1995) Speech recognition with primarily temporal cues. Science 270:303–304.

Sohoglu E, Davis MH (2016) Perceptual learning of degraded speech by minimizing prediction error. Proceedings of the National Academy of Sciences of the United States of America 113:E1747-56.

Sohoglu E, Peelle JE, Carlyon RP, Davis MH (2012) Predictive Top-Down Integration of Prior Knowledge during Speech Perception. The Journal of neuroscience : the official journal of the Society for Neuroscience 32:8443–8453.

Sohoglu E, Peelle JE, Carlyon RP, Davis MH (2014) Top-Down Influences of Written Text on Perceived Clarity of Degraded Speech. J Exp Psychol Human 40:186–199.

Summerfield C, Trittschuh EH, Monti JM, Mesulam MM, Egner T (2008) Neural repetition suppression reflects fulfilled perceptual expectations. Nat Neurosci 11:1004–1006.

Teufel C, Subramaniam N, Dobler V, Perez J, Finnemann J, Mehta PR, Goodyer IM, Fletcher PC (2015) Shift toward prior knowledge confers a perceptual advantage in early

psychosis and psychosis-prone healthy individuals. Proceedings of the National Academy of Sciences 112:13401–13406.

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M (2002) Automated Anatomical Labeling of Activations in SPM Using a Macroscopic Anatomical Parcellation of the MNI MRI Single-Subject Brain. NeuroImage 15:273–289.

Zoefel B, Davis MH (2017) Transcranial electric stimulation for the investigation of speech perception and comprehension. Language, Cognition and Neuroscience 32:910–923.

Figures and Legends

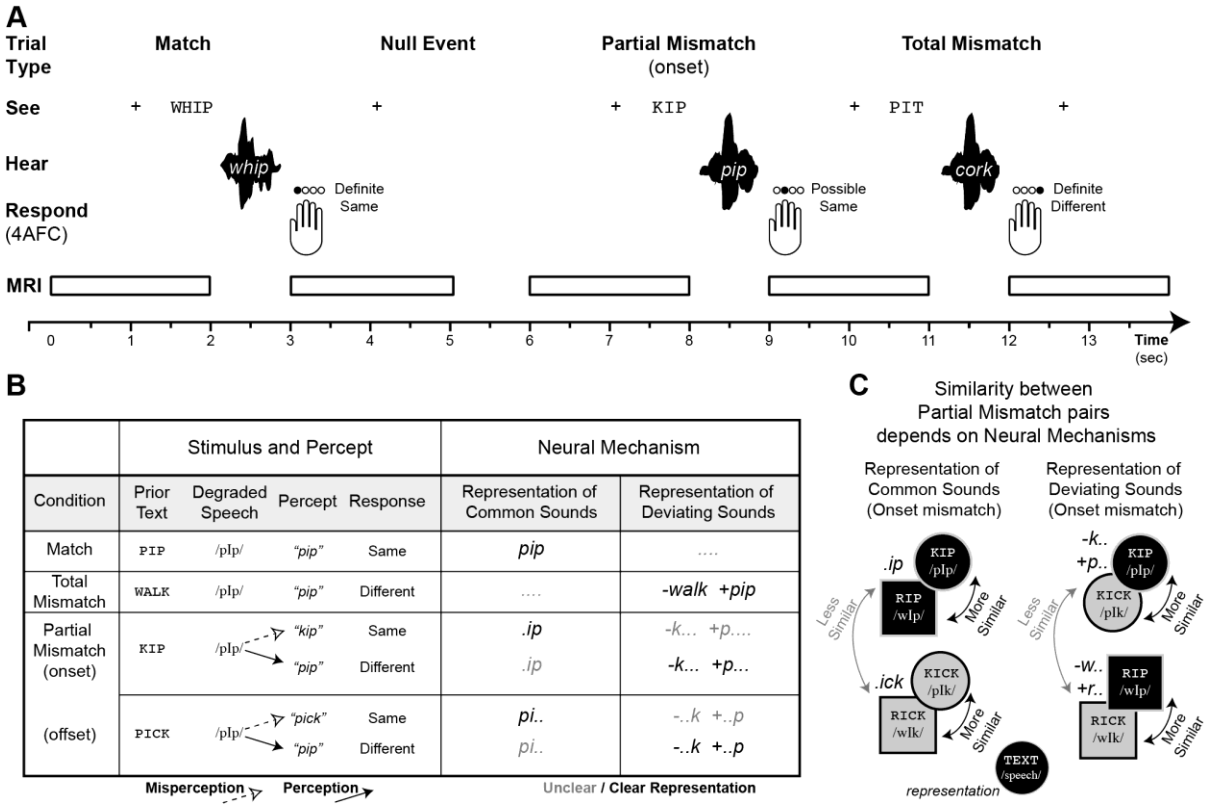

Figure 1

Figure 1. Experimental design and hypotheses. A. Experimental design. We used fMRI to measure brain activity while participants read written words and heard subsequent degraded spoken words. Written and spoken words were combined in three conditions: 1) Match (identical written/spoken words, e.g., whip–whip), 2) Partial Mismatch (e.g., kip–pip, or pick-pip), and 3) Total Mismatch (e.g., pit–corn). Participants responded with a button press to indicate whether spoken/written words were **"**same**"** or **"**different**"** and their confidence. B. Stimulus conditions, responses, and underlying neural mechanisms for representing written/spoken word pairs. In an Onset Partial Mismatch trial (depicted in the third row) the spoken word /pIp/ ("pip") following written KIP can be perceived as *"kip"* (= "same" response) or *"pip"* ("different"). This behavioural outcome could be explained by one of two neural mechanisms. 1. A representation of common sounds would produce a clear

representation of the sounds "*.ip"* (shown in black); this representation would be clearer (black) for trials in which participants report that written/spoken words are the "same" than if participants report that written/spoken words differ. 2. A representation of deviating sounds would produce a clearer representation of the deviating sounds -*..k* +*..p* (black) on trials in which participants report that written/spoken words differ, and an unclear representation of the deviating sounds -*..k* +*..p* (light grey) if participants report that written/spoken words are the "same". C. Similarity between Partial Mismatch pairs depends on the underlying neural mechanism. A neural representation of common sounds in written/spoken word pairs predicts that representations for word pairs sharing the same expected sounds (grouped by colour, left side) should be more similar (e.g., KIP-/pIp/ share sounds *.ip* and are therefore more similar to RIP-/wIp/, also sharing *.ip*, than to KICK-/pIk/ or RICK-/wIk/ sharing *.ick*). In contrast, a neural representation of deviating sounds (i.e., prediction error, grouping by shape, right side), predicts that word pairs sharing the same deviating sounds should be more similar (e.g., KIP-/pIp/ deviate in *–k…+p* and should be more similar to KICK-/pIk/, also deviating in *–k…+p*, than to RIP-/wIp/ and RICK-/wIk/ deviating in *–r…+w*). Similar examples apply in other conditions (i.e., Offset Partial Mismatch trials), ensuring that differential representation of onset and offset segments does not favour one or other account.

Figure 2

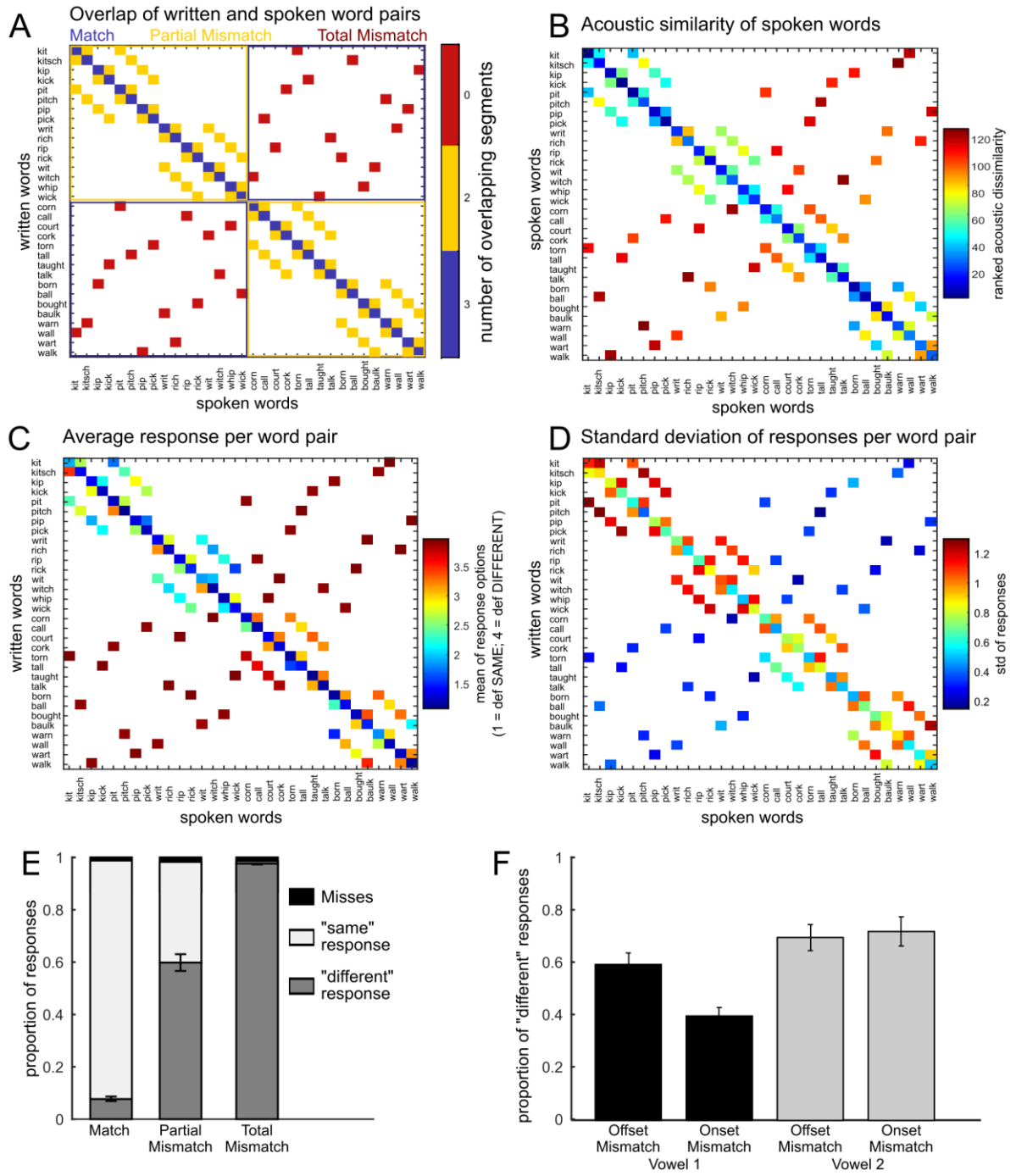

Figure 2. Stimulus Similarity, Behavioural Confusion Matrix, and Behavioural Results.

A. Stimulus Similarity Matrix. We combined 32 written words with 32 spoken words in three different conditions (Match, Total Mismatch, and Partial Mismatch at Onset or Offset), so that the experiment contained 128 different spoken/written word pairs. Pairs of written and spoken

words had varying numbers of overlapping sounds in three different experimental conditions. In Match trials, all three speech segments overlapped (blue diagonal), Total Mismatch trials, no segments overlapped (red). In Partial Mismatch trials, two segments overlapped between written and spoken words (yellow). **B. Acoustic Similarity.** Acoustic dissimilarity between 6-channel vocoded spoken words was computed using methods described by (Billig et al., 2013) shown for critical word pairs in rank order. **C. Mean behavioural responses.** Participants responded to each word pair to indicate whether written and spoken words matched and their confidence (1="definitely same", 2="possibly same", 3="possibly different", 4="definitely different"). Match trials were perceived as "definitely same". Total Mismatch trials were perceived as "definitely different". Partial Mismatch trials were perceived as "same" or "different" with reduced confidence. **D. Standard deviation of responses per word pair.** Behavioural responses in the Match and Total Mismatch condition were consistent (blue), whereas responses in the Partial Mismatch condition were more variable. **E.** Behavioural responses showed more "same" responses (light grey bars) in the Match than in Partial Mismatch and Total Mismatch conditions. Conversely, participants responded correctly ("different") in a large proportion of Partial Mismatch and in almost all Total Mismatch trials (dark grey bars). Error bars show standard error of the mean over subjects corrected for repeated measures comparisons. **F.** Proportion of "different" responses shown separately for Partial Mismatch trials conditions split by Vowel and for Onset/Offset Partial Mismatch. Error bars show the standard error of the mean over items.

Figure 3

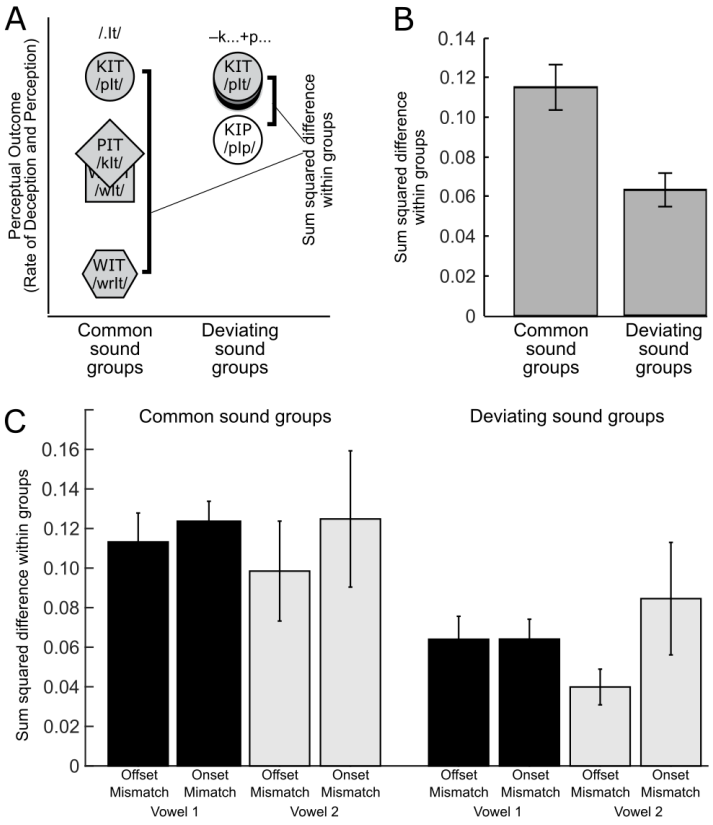

Figure 3. Perception of Partial Mismatch pairs is predicted by the identity of deviating sounds A. Schematic illustration of behavioural analysis of Partial Mismatch trials. For each Partial Mismatch pair, we computed the sum square difference between the rate of "different" responses for that item pair and the three other Partial Mismatch word pairs that share either the same (1) common sounds (e.g. kit-pit compared with pit-kit, wit-writ, etc.) or (2) deviating sounds (e.g. kit-pit compared with kip-pip, kitsch-pitch, etc.). This analysis is independent of the overall rate of "different" responses but considers the consistency of responses between items within the same group. **B.** Perceptual outcomes were significantly more similar for word pairs sharing the same deviating sounds than for word pairs sharing the same common sounds (i.e., reduced sum squared difference). Error bars show the standard error of the mean over items. **C.** Mean Squared Differences for common and deviating Sound groups split by Vowel and for Onset/Offset Partial Mismatch.

Figure 4

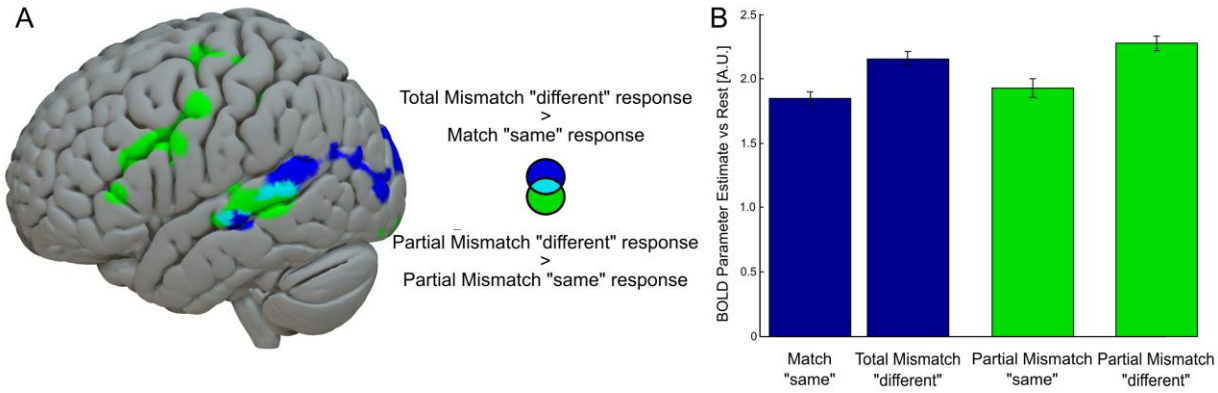

Figure 4. Univariate fMRI results A. Whole brain fMRI analysis showed overlapping response increases in the left STS for two key contrasts: Total Mismatch ("different" response) > Match ("same" response) (blue) and Partial Mismatch ("different" response) > Partial Mismatch ("same" response) (green). Overlapping responses are shown in cyan (both contrasts are displayed at $p < 0.001$, uncorrected but reach $p < 0.05$ FWE cluster-corrected significance in left STS; see Tables 2 and 3). **B.** BOLD parameter estimates vs rest in the left posterior STS extracted from the overlapping region activated for the two contrasts: Total Mismatch ("different" response) > Match ("same" response) and Partial Mismatch ("different" response) > Partial Mismatch ("same" response). Error bars show the standard error of the mean over participants after between participant variance is removed, suitable for repeated measures comparisons.

Figure 5

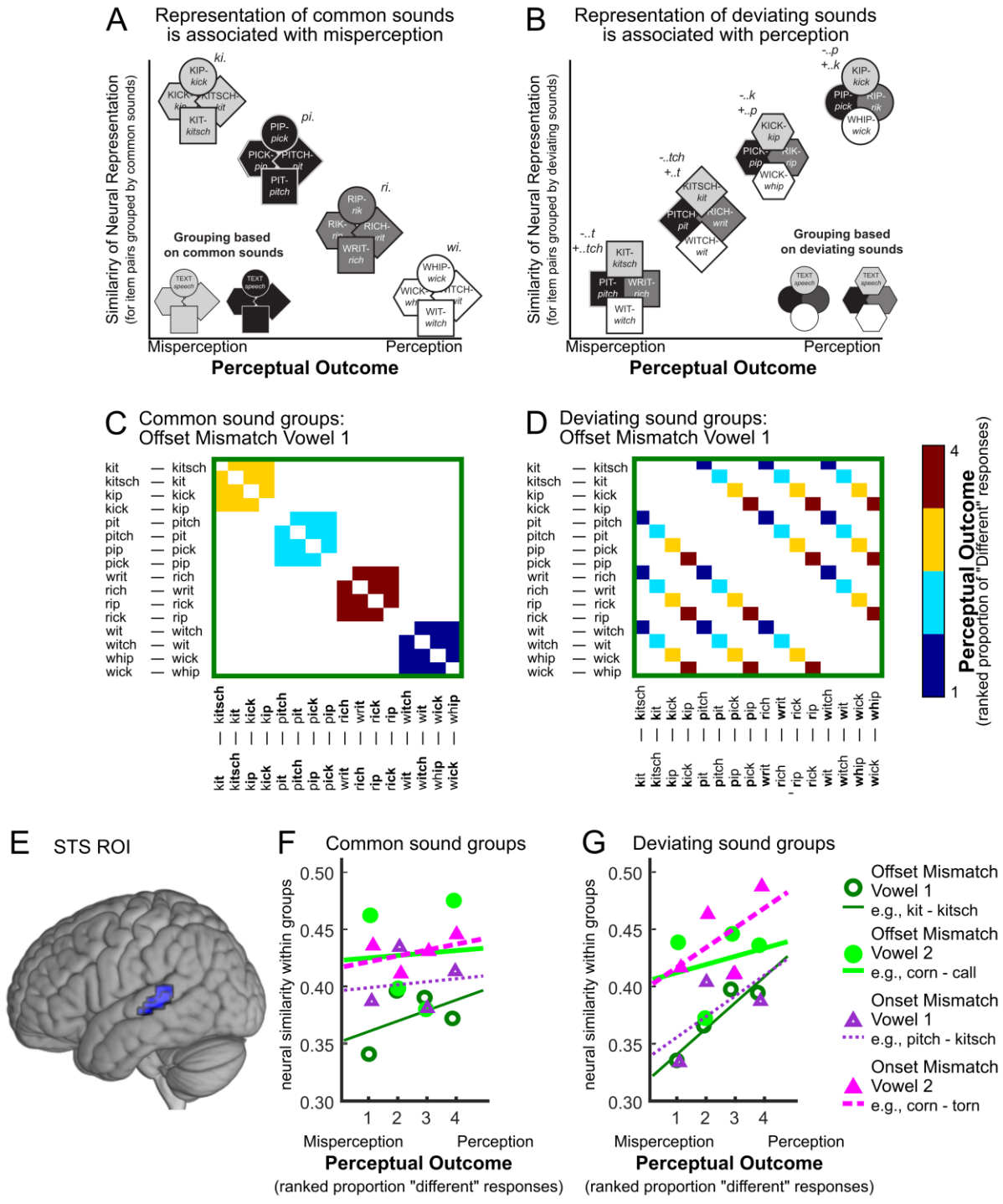

Figure 5. Representational Similarity Analysis Predictions, Methods and pSTS Results. A. Predicted correlation if misperception is associated with representations of common sounds in Partial Mismatch trials. If neural representations preferentially represent common sounds in written/spoken word pairs then Partial Mismatch pairs that share the same common

sounds (e.g., kit-kick, kit-kitsch, etc.) should generate similar neural representations (groups of items with the same common sounds are indicated by the same colour). Clearer representations of common sounds (i.e., increased neural similarity within groups) should lead to confirmation of the prior (i.e., "same" responses, misperception) while less clear representations lead to rejection of the prior ("different" responses, correct perception). Hence, representation of common sounds in Partial Mismatch trials predicts a negative correlation between neural similarity and perception. **B. Predicted correlation if perception is associated with representations of deviating sounds in Partial Mismatch trials.** If neural representations preferentially represent deviating sounds in written/spoken word pairs then Partial Mismatch pairs that share the same deviating sounds (e.g., ki**p**-ki**ck**, whi**p**-wi**ck**, etc.) should generate similar neural representations (groups of items with the same deviating sounds are indicated by the same shape). Clearer representations of these deviating sounds (i.e., increased neural similarity within groups) should lead to rejection of the prior (i.e., "different" responses, correct perception) while less clear representations of deviating sounds lead to confirmation of the prior ("same" responses, misperception). Hence, representation of deviating sounds in Partial Mismatch trials predicts a positive correlation between neural similarity and perception. Hypothesis RDMs for comparisons of word pairs that **(C)** shared the same common sounds in written/spoken word pairs (e.g., for Offset Mismatch pairs containing the vowel /I/: **ki**t-**ki**tsch, **ki**ck-**ki**p; here the common sounds /**kI**/ are the same for these word pairs), **or D.** shared the same deviating sounds in written/spoken word pairs (e.g., for Offset Mismatch pairs containing the vowel 1: **ki**t-**ki**tsch, pi**t**-pi**tch**; despite the different spellings these contain the same deviating /**t**/ and /**t**ʃ/ sounds). In other hypothesis RDMs (not shown) we applied the same principle for Onset Mismatch pairs like pick-kick and for the vowel /ɔ:/ as in tall-call. We can supply the other RDMs for interested readers on request. For visualization, we show a

hypothesis RDM based on the average ranking of “different” responses across participants; for analysis different rankings were used based on behavioural data from individual participants. **E.** The search volume in the left STS used in these analyses was defined from an independent univariate contrast Total Mismatch (“different” response) > Match (“same” response) (see Methods and Fig 3A) masked to confine analysis to the superior temporal sulcus. (**F/G). Correlation of neural similarity and perceptual outcomes.** Results are visualized for 16 data points for four different sets of word pair groups and the factorial crossing of Offset/Onset Partial Mismatch word pairs containing the two vowels /I/ and /ɔː/ as in *kick*, and *tall*. Lines show the least-square fit to the data. F**. Common sound groups.** When Partial Mismatch trials were grouped by common sounds, neural similarity did not correlate with perception as hypothesized (compare with panel A). **G. Deviating sound groups.** When Partial Mismatch trials are grouped by deviating sounds neural similarity correlated positively with perception as hypothesized (see panel B). Results and least-square fit lines as in F.

Figure 6

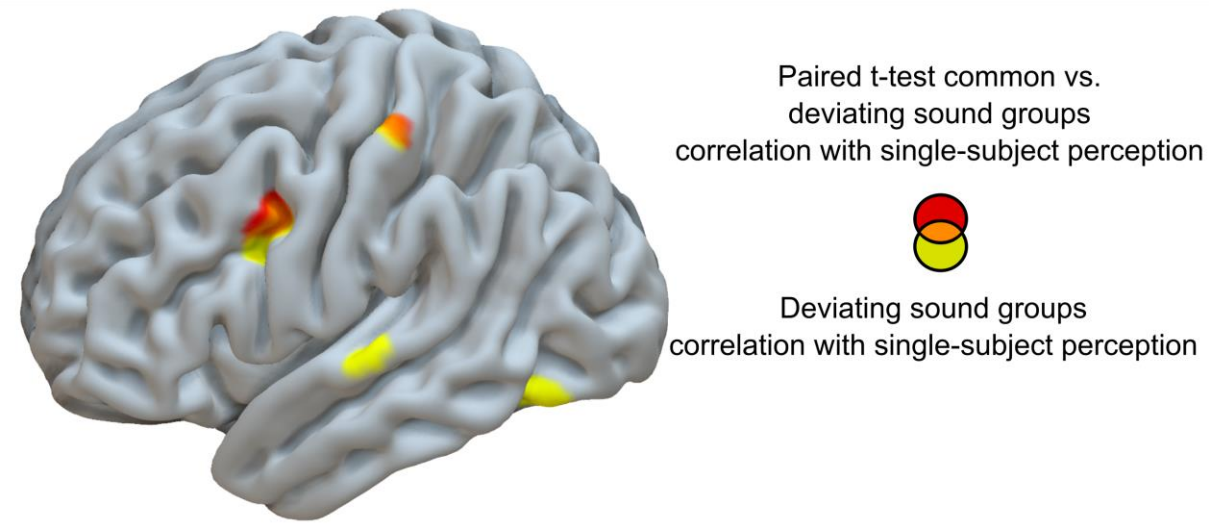

Figure 6. Results of the whole-brain searchlight RSA approach (shown at $p < 0.001$, uncorrected for clarity). The paired t-test comparing the correlation between single-subject perception and neural representations of common vs. deviating sounds is shown in red. The correlation between single-subject perception and representations of deviating sounds is shown in yellow. Corresponding coordinates are reported in Tables 4 and 5.

Tables

Table 1 Partial Mismatch pairs

		Onset Partial Mismatch				Offset Partial Mismatch			
word pair	vowel number	written word	spoken word	deviating sound	common sound	written word	spoken word	deviating sound	common sound
1	1	kit	pit	k/p	It	kit	kitsch	t/tʃ	kI
2	1	kitsch	pitch	k/p	Itʃ	kitsch	kit	tʃ/t	kI
3	1	kip	pip	k/p	Ip	kip	kick	p/k	kI
4	1	kick	pick	k/p	Ik	kick	kip	k/p	kI
5	1	pit	kit	p/k	It	pit	pitch	t/tʃ	pi
6	1	pitch	kitsch	p/k	Itʃ	pitch	pit	tʃ/t	pi
7	1	pip	kip	p/k	Ip	pip	pick	p/k	pi
8	1	pick	kick	p/k	Ik	pick	pip	k/p	pi
9	1	writ	wit	r/w	It	writ	rich	t/tʃ	ri
10	1	rich	witch	r/w	Itʃ	rich	writ	tʃ/t	ri
11	1	rip	whip	r/w	Ip	rip	rick	p/k	ri
12	1	rick	wick	r/w	Ik	rick	rip	k/p	ri
13	1	wit	writ	w/r	It	wit	witch	t/tʃ	wi
14	1	witch	rich	w/r	Itʃ	witch	wit	tʃ/t	wi
15	1	whip	rip	w/r	Ip	whip	wick	p/k	wi
16	1	wick	rick	w/r	Ik	wick	whip	k/p	wi
17	2	corn	torn	k/t	ɔːn	corn	call	n/l	kɔː
18	2	call	tall	k/t	ɔːl	call	corn	l/n	kɔː
19	2	court	taught	k/t	ɔːt	court	cork	t/k	kɔː
20	2	cork	talk	k/t	ɔːk	cork	court	k/t	kɔː
21	2	torn	corn	t/k	ɔːn	torn	tall	n/l	tɔː
22	2	tall	call	t/k	ɔːl	tall	torn	l/n	tɔː
23	2	taught	court	t/k	ɔːt	taught	talk	t/k	tɔː
24	2	talk	cork	t/k	ɔːk	talk	taught	k/t	tɔː

25	2	born	warn	b/w	ɔːn	born	ball	n/l	bɔː
26	2	ball	wall	b/w	ɔːl	ball	born	l/n	bɔː
27	2	bought	wart	b/w	ɔːt	bought	balk	t/k	bɔː
28	2	balk	walk	b/w	ɔːk	balk	bought	k/t	bɔː
29	2	warn	born	w/b	ɔːn	warn	wall	n/l	wɔː
30	2	wall	ball	w/b	ɔːl	wall	warn	l/n	wɔː
31	2	wart	bought	w/b	ɔːt	wart	walk	t/k	wɔː
32	2	walk	balk	w/b	ɔːk	walk	wart	k/t	wɔː

Table 1 Grey color in the deviating and common sound columns indicates the group number.

Colours for the written/spoken word pairs is as shown in Fig 4

Table 2 Univariate fMRI Analysis: Total Mismatch "different" percept > Match "same" percept, displayed at p < 0.001 uncorrected and more than 10 voxels per cluster. Brain regions are labelled based on the AAL atlas (Tzourio-Mazoyer et al., 2002).

cluster p(FWE-corr)	Cluster size	peak p(FWE-corr)	peak equivZ	x,y,z [mm]	Anatomical label of the peak
0.094	116	0.035	4.94	-38 -28 60	Left postcentral gyrus
0.000	758	0.045	4.88	-66 -36 14	Left superior temporal gyrus
		0.257	4.39	-58 -34 6	Left middle temporal gyrus
		0.339	4.30	-58 -62 24	Left angular gyrus
0.002	291	0.088	4.70	-32 20 32	Left middle frontal gyrus
		0.705	3.98	-38 16 22	Left inferior frontal gyrus, pars opercularis
0.000	548	0.112	4.64	-10 -98 12	Left superior occipital
		0.150	4.55	-8 -92 20	Left cuneus
		0.689	3.99	-20 -94 24	Left superior occipital
0.006	231	0.176	4.51	60 -8 -10	Right middle temporal gyrus
		0.376	4.26	66 -16 -6	Right middle temporal gyrus
		0.996	3.46	68 -22 2	Right superior temporal gyrus
0.005	232	0.511	4.14	-4 -48 54	Left precuneus
		0.941	3.70	-10 -54 36	Left precuneus
		0.997	3.45	-2 -54 46	Left precuneus
0.434	60	0.612	4.06	-26 -58 -12	Left fusiform gyrus
0.629	45	0.649	4.03	12 50 -6	Right medial orbitofrontal cortex
0.004	246	0.673	4.01	12 -68 -4	Right lingual gyrus
		0.885	3.79	18 -76 -4	Right lingual gyrus
		0.983	3.57	20 -70 -10	Right lingual gyrus
0.298	74	0.816	3.87	-42 -88 6	Left middle occipital gyrus
		0.987	3.55	-36 -92 10	Left middle occipital gyrus
		1.000	3.27	-46 -80 14	Left middle occipital gyrus
0.001	320	0.842	3.85	42 -44 16	Right middle temporal gyrus

		0.880	3.80	62 -50 24	Right superior temporal gyrus
		0.943	3.70	46 -68 32	Right angular gyrus
0.716	39	0.952	3.68	-4 -22 48	Left midcingulate area
		1.000	3.19	-8 -12 50	Left supplementary motor area
0.788		0.969	3.63	22 -46 -22	Right lobule IV, V of cerebellar
	34				hemisphere
0.987	13	0.999	3.35	-16 -48 -8	Left lingual gyrus

Table 3 Univariate fMRI Analysis: Partial Mismatch "different" percept > Partial Mismatch "same" percept, displayed at p < 0.001 uncorrected and more than 10 voxels per cluster. Brain regions are labelled based on the AAL atlas.

cluster p(FWE-corr)	Cluster size	peak p(FWE-corr)	peak equivZ	x,y,z [mm]	Anatomical label of the peak
0.001	374	0.005	5.35	-44 -54 -16	Left fusiform gyrus
		0.340	4.22	-40 -46 -18	Left fusiform gyrus
		0.923	3.64	-36 -40 -24	Left fusiform gyrus
0.000	662	0.022	4.98	22 -50 -30	Right lobule VI of cerebellar hemisphere
		0.035	4.87	16 -48 -24	Right lobule IV, V of cerebellar hemisphere
		0.411	4.15	26 -60 -34	Right lobule VI of cerebellar hemisphere
0.000	1146	0.029	4.92	-50 26 18	Left inferior frontal gyrus, pars triangularis
		0.353	4.21	-42 12 34	Left precentral gyrus
		0.504	4.07	-36 22 18	Left inferior frontal gyrus, pars triangularis
0.005	297	0.058	4.74	-20 -4 60	Left superior frontal gyrus
		0.310	4.25	-20 -6 52	Left superior frontal gyrus
		1.000	3.23	-26 6 60	Left middle frontal gyrus
0.442	69	0.201	4.39	6 0 -6	Right globus pallidus
0.000	658	0.312	4.25	-54 -28 -2	Left middle temporal gyrus
		0.518	4.05	-62 -30 2	Left middle temporal gyrus
		0.804	3.80	-50 -42 6	Left middle temporal gyrus
0.238	98	0.449	4.11	-32 -26 56	Left precentral gyrus
		0.991	3.42	-42 -30 58	Left postcentral gyrus
0.068	157	0.486	4.08	-18 -38 -20	Left lobule IV, V of cerebellar hemisphere
0.271	92	0.697	3.90	-28 -6 -2	Left putamen
0.986	12	0.808	3.79	-32 -70 32	Left middle occipital gyrus

0.562	57	0.819	3.78	-10 32 48	Left medial frontal gyrus
0.233	99	0.939	3.61	-14 -52 -28	Left lobule IX of cerebellar hemisphere
		0.951	3.59	-24 -48 -30	Left lobule VI of cerebellar hemisphere
0.989	11	0.981	3.49	4 58 -14	Right gyrus rectus
0.961	18	0.982	3.48	12 -58 -46	Right lobule IX of cerebellar hemisphere
0.986	12	0.987	3.46	-10 -28 -14	Left lobule III of cerebellar hemisphere
0.989		0.998	3.32	14 -82 -34	Right crus II of cerebellar hemisphere"
	11				0
0.986	12	0.998	3.30	-12 -94 10	Left superior occipital
0.999	4	0.998	3.30	-6 -62 24	Left cuneus
0.986	12	0.999	3.29	2 -62 10	Right Calcarine sulcus 0

Table 4 RSA fMRI Analysis: paired t-test comparing the correlation between single-subject perception and representations of common vs. deviating sounds, reported at p < 0.001 uncorrected. Brain regions are labelled based on the AAL atlas.

cluster p(FWE-corr)	Cluster size	peak p(FWE-corr)	peak T	peak equivZ	x,y,z [mm]	Anatomical label of the peak
0.114	148	0.273	4.37	3.69	-30 2 39	Left middle frontal gyrus
		0.603	3.78	3.30	-54 20 39	Left middle frontal gyrus
0.366	49	0.356	4.19	3.58	-45 -28 58	Left postcentral gyrus

Table 5 RSA fMRI Analysis: Correlation between single-subject perception and representations of deviating sounds, reported at p < 0.001 uncorrected. Brain regions are labelled based on the AAL atlas.

cluster p(FWE-corr)	Cluster size	peak p(FWE-corr)	peak T	peak equivZ	x,y,z [mm]	Anatomical label of the peak
0.267	73	0.212	4.55	3.81	-66 -16 9	Left superior temporal gyrus
0.050	220	0.215	4.55	3.80	-30 2 39	Left middle frontal gyrus
		0.505	3.97	3.43	-33 5 20	Left insula
0.160	114	0.287	4.37	3.69	-45 -67 -14	Left fusiform gyrus
0.330	57	0.407	4.13	3.54	-45 -22 65	Left postcentral gyrus
0.453	34	0.525	3.94	3.41	54 -7 -2	Right superior temporal gyrus
0.698	6	0.628	3.78	3.30	-33 -1 -25	Left hippocampus
0.740	3	0.773	3.55	3.13	-42 -61 24	Left angular gyrus